Université Jean Monnet

Data Mining

Project Report

Meteorite Analysis

Author:
Omar Elsabrout

Supervisor:
Dr. Fabrice Muhlenbach

March 23, 2018

Abstract

Through out the years, a huge number of meteorites has fallen to Earth from outer space. These incidents are recorded by the Meteoritical Society and stored in a dataset that includes the location, mass, composition, and fall year for over 45,000 meteorites that have struck our planet. Such an interesting topic is perfect to be utilized as a study example for data mining techniques we try to learn in our Data Mining project.

1 Introduction

Since the objective of the project is to search a real life problem and to find a sufficiently large dataset for being able to apply data mining techniques with **R** to find some answers to this problem or to extract new strange interesting knowledge out of it.

A relatively large dataset is selected for the project to find interesting, unexpected or valuable structures that are embedded in its data. This dataset was downloaded from NASA's Data Portal, and is based on The Meteoritical Society's Meteoritical Bulletin Database (this latter database provides additional information such as meteorite images, links to primary sources, etc.). The dataset can be downloaded from the following link (`https://data.nasa.gov/Space-Science/Meteorite-Landings/ak9y-cwf9`)

2 Problem Understanding

Concerning the given dataset of meteorite landings, it does not have the regular paradigm of a dataset designed for problem solving. On the other hand, it represents observations of nature and outer space which makes the dataset belong to the category which we only seek to extract some phenomenon or strong correlations. Until now, space still hold several unexplained phenomena and the scientific community has been collecting data for many years to help us mine this knowledge and reach new conclusions.

Getting into understanding what meteorites are, a meteorite is a solid piece of debris from an object, such as a comet, asteroid, or meteoroid, that originates in outer space and survives its passage through the atmosphere to reach the surface of a planet or moon. Meteorites that are recovered after being observed as they transit the atmosphere or impact the Earth

are called meteorite falls. All others are known as meteorite finds. As of April 2016, there were about 1,140 witnessed falls that have specimens in the world's collections. As of 2011, there are more than 38,660 well-documented meteorite finds.

One of the features tackled in this report is the type of meteorite recovery. Hence, it is crucial to understand the types of recoveries of meteorites before looking into the dataset. There are two meteorite recovery types which are Falls and Finds. Most meteorite falls are recovered on the basis of eyewitness accounts of the fireball or the impact of the object on the ground, or both. Therefore, despite the fact that meteorites fall with virtually equal probability everywhere on Earth, verified meteorite falls tend to be concentrated in areas with high human population densities such as Europe, Japan, and northern India. A small number of meteorite falls have been observed with automated cameras and recovered following calculation of the impact point. NASA, which is the provider of our dataset, has an automated system that detects meteors and calculates the orbit, magnitude, ground track, and other parameters over the southeast USA, which often detects a number of events each night.

3 Data Understanding

One of the most important concerns when it comes to data mining in general is understanding the dataset and its purpose before starting to process it. In our case, we have to understand what are meteorites and their landings which is explained in the previous part. On the other hand, in this section we go more into details regarding the dataset itself. Our dataset is imported from NASA's open data portal in a CSV format to be imported to R studio. Our dataset consists of ten features which are 'name', 'id', 'nametype', 'recclass', 'mass', 'fall', 'year', 'reclat', 'reclong' and 'GeoLocation' respectively. The dataset has 45,617 rows accordingly which makes it a great setup for unsupervised learning since we have no labels for each row.

In order to show our work on the data, a specific explanation for every column has to be presented to understand the purpose of our analysis. In a column by column manner, we start with the most basic feature which is the 'name' which basically expresses the name of the meteorite in a string form. Following, the 'id' feature which is surprisingly does not depend on the time of discovery of the meteorite and seems to be randomly selected. Nonethe-

less, it is a valid identification of the meteorite since no two rows have the same 'id'. Then, the 'nametype' feature which has the value "Valid" for all meteorites which is completely useless. However, we might think it is a left-over of another dataset that contained valid and invalid rows and then it got filtered. Next, we have the 'recclass' feature which is a text that represents the elements forming this meteorite. This information is valuable from a geological and astrophysical point of view and is helpful to our understanding of the different types of meteorites. Then, one of the most important features is 'mass' since it expresses the mass of the meteorite in kilograms. Next, we check the feature 'fall' which has only two values either Found or Fall. This indicates the type of meteorite previously explained which contributes significantly to our understanding of the data. Then, the feature 'year' shows the time of the landing of the meteorite which is an extremely difficult task if the meteorite is already found on earth. Hence, there are several rows without a value for this feature which is understandable but causes to lower the quality of the data. The following two features are correlated which are 'reclat' and 'reclong'. They represent the location coordinates of the meteorite which takes us to the following feature. Finally, the last feature is 'GeoLocation' which is redundant since it depends completely and only on the 'reclat' and 'reclong' features without any addition of information.

This concludes the explanation of all of our features and their usefulness. Our next step is to prepare these features and adjust the data in general to start our analysis. Such task is crucial and encouraged as it optimizes the analysis process.

4 Data Preparation

Most of available dataset are collected only for the purpose of preserving knowledge. However, that does not mean they are ready for knowledge mining instantly. Several observations and manipulations are in order to extract this knowledge and analyze it.

In our case, some modifications are done to make the data as expressive as possible. First, we discarded the feature 'GeoLocation' since it adds absolutely no new knowledge and it is composed of the 'reclat' and 'reclong' features. Secondly, we reformat the feature 'year' as the rows do not have the same format. Some of them are dates with days, months and years. On the other hand, some of them are only years. We decided that only years are

relevant in our analysis. Such process can be done using R in a script but we found that it can be done just as easily in Microsoft Excel without the need to implement a script or code snippet. At the end, only results matter and efficiency dictates that we select the simplest method if we do not sacrifice quality.

At last, we realize that the data is incomplete in several rows due to the fact that it is collected from several entities with different quality measures. Hence, we filter incomplete rows before getting into the analysis process. By filtering we mean removing rows with missing features, rows in invalid years and rows with strange and out of the scope locations. This helps us in reaching solid and reliable results and removes noise from the data.

5 Modeling and Evaluation

In our report, we present the mined knowledge in form of sections since modeling in our case is unconventional as a result of observing natural phenomena and not a dedicated dataset to solve a specific problem. The paradigm of our project is different from the conventional form because we had to adapt our prospective to match our dataset.

5.1 Class Analysis

5.1.1 Meteorite Class Count

We plot the twenty most occurring meteorites in a flipped bar plot in Figure 1. This indicates which classes are the most frequent.

5.1.2 Meteorite Mass Count

We plot the twenty most heavy meteorites based on their median mass in a flipped bar plot. in Figure 2. This shows the correlation between meteorite mass and frequency.

5.2 Distribution of Mass

Here Mass is a continous variable and therfore for the distribution we plot a histogram. We plot the distribution of the Mass of the meteorites in Figure 3.

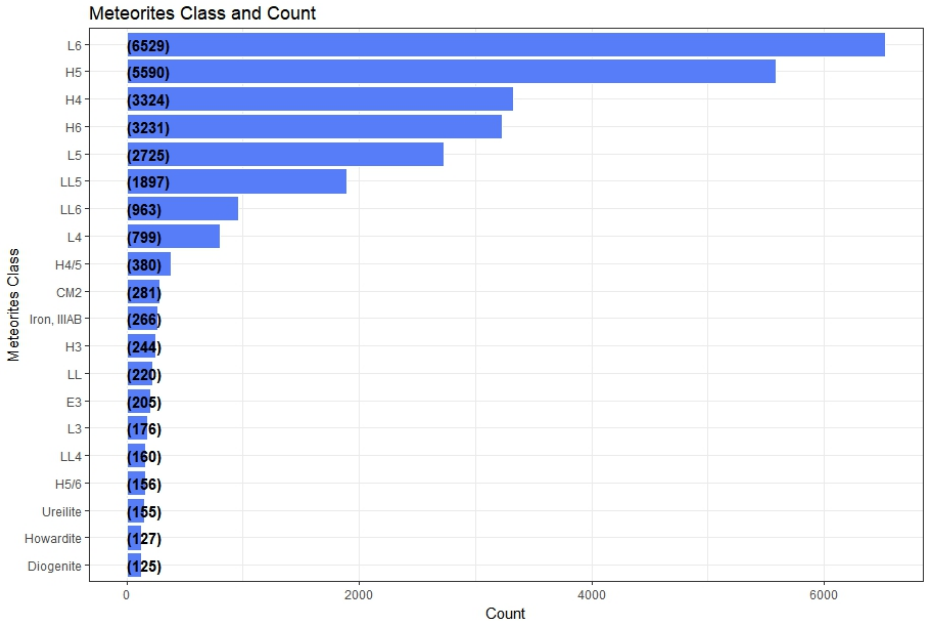

Figure 1: Meteorites Class and Count.

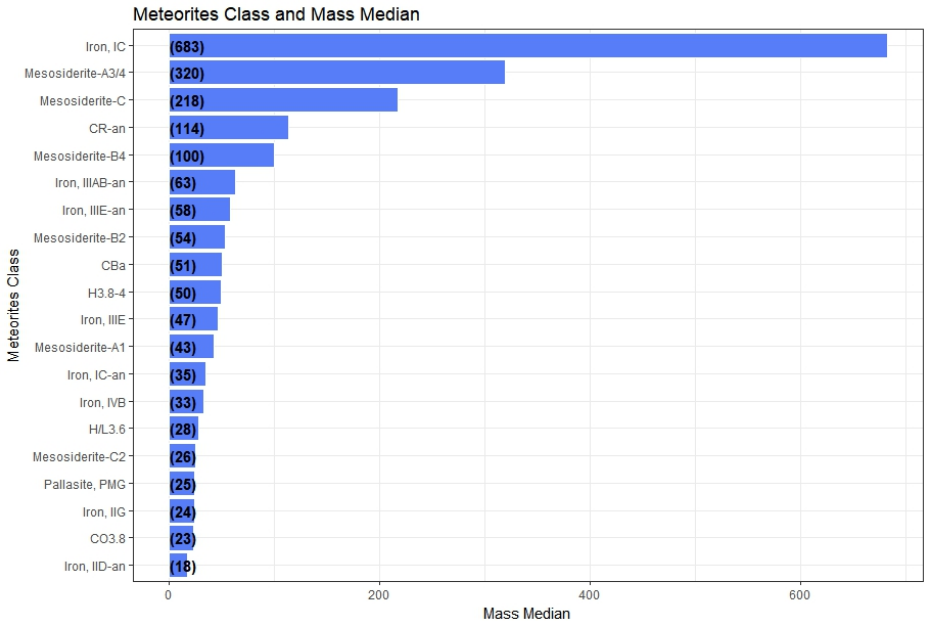

Figure 2: Meteorites Mass and Count.

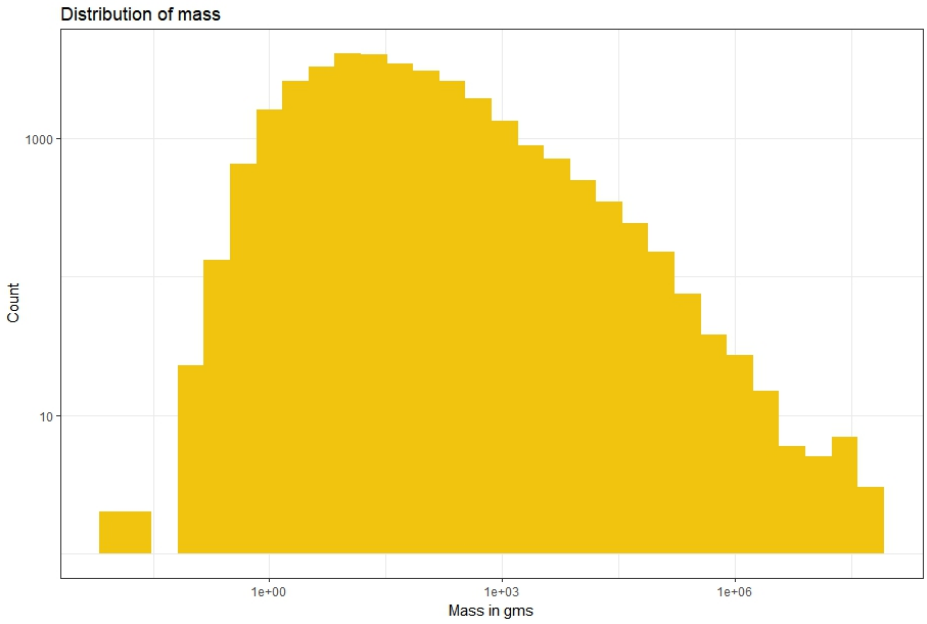

Figure 3: Distribution of Mass.

5.2.1 Heaviest Meteorite

We extract the heaviest meteorite with its mass is in Kilograms.

name	id	nametype	recclass	mass	fall	year	reclat	reclong	GeoLocation
Hoba	11890	Valid	Iron, IVB	60000	Found	1920	-19.58333	17.91667	(-19.583330, 17.916670)

5.2.2 Map of the Heaviest Meteorite

The mass of the meteorites are indicated by the radius of the circles. The radius is equal to the mass in kilograms multiplied by 10 in Figure 4.

5.2.3 Lightest Meteorite

We extract the lightest meteorite with its mass is in Kilograms.

name	id	nametype	recclass	mass	fall	year	reclat	reclong	GeoLocation
Yamato 8333	29438	Valid	H5	0.00001	Found	1983	-71.5	35.66667	(-71.500000, 35.666670)

Figure 4: Heaviest Meteorite.

5.2.4 Distribution of Mass classified by Fall Type

We plot the distribution of the mass of the meteorites based on their fall type in Figure 5. To examine the relationships between a continuous and categorical variable, we plot a facet bar plot in Figure 6.

5.3 Distribution of Meteorite Landings with Meteorite Mass

The following plot shows the distribution of the meteorite landings all over the world. The mass of the Meteorites are indicated by the Radius of the Circles. Plot is shown in Figure 7.

5.4 Distribution of US Meteorite Landings

The following plot shows the distribution of the US meteorite landings.Here we have filtered the US meteorite landings by filtering the latitude and longitude in Figure 8. The heaviest ten of them are represented in the following table:

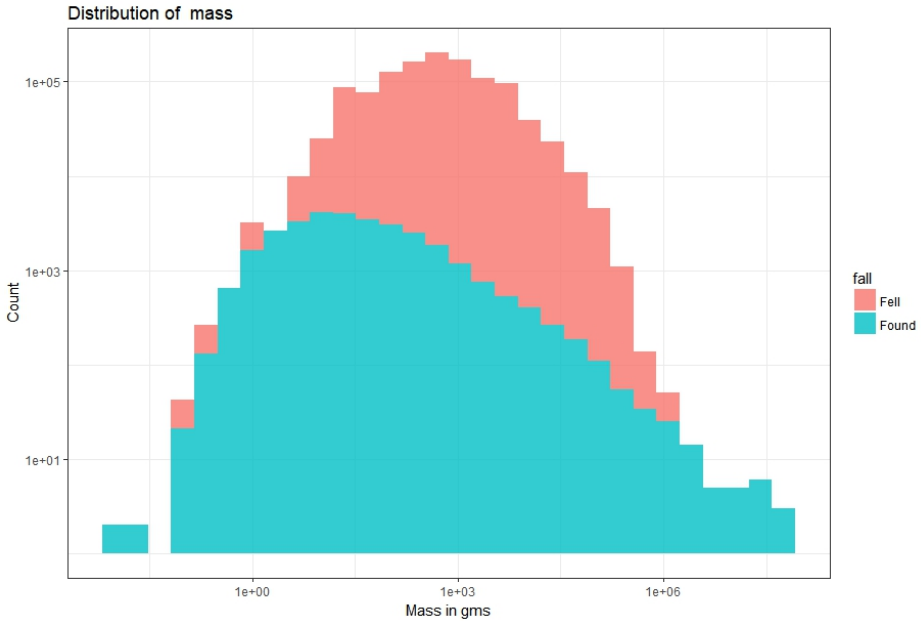

Figure 5: Distribution Of Mass.

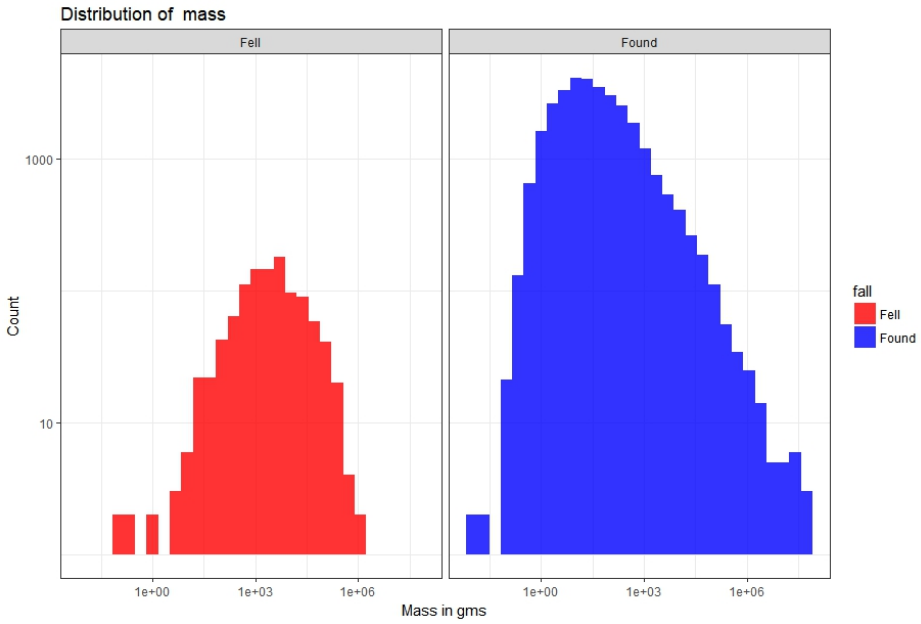

Figure 6: Distribution Of Mass.

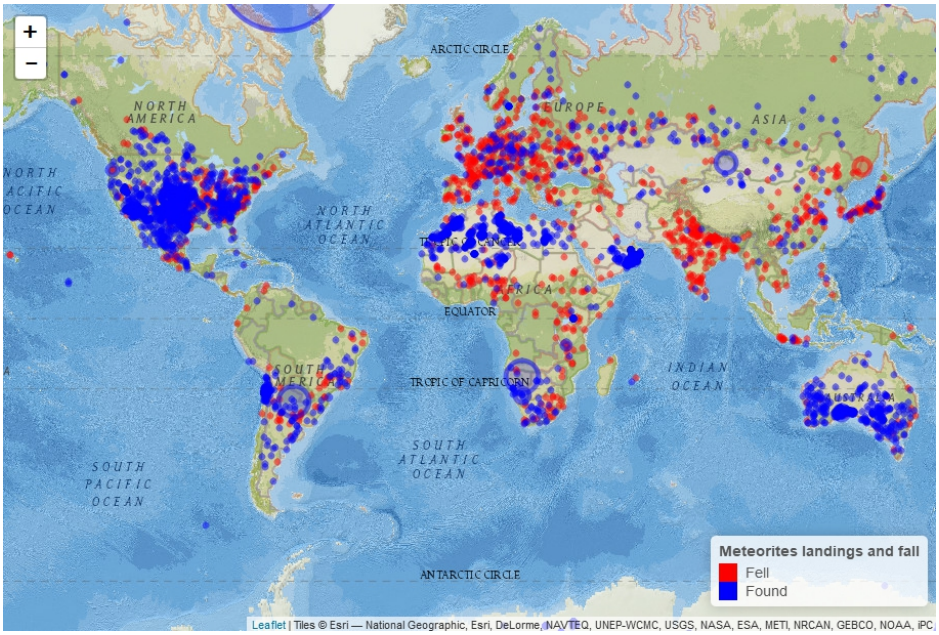

Figure 7: Map of Meteorites.

	name	id	nametype	recclass	mass	fall	year	reclat	reclong	GeoLocation
1	Canyon Diablo	5257	Valid	Iron, IAB-MG	30000000	Found	1891	35.05	-111.03333	(35.050000, -111.033330)
2	Chupaderos	5363	Valid	Iron, IIIAB	24300000	Found	1852	27	-105.1	(27.000000, -105.100000)
3	Bacubirito	4919	Valid	Iron, ungrouped	22000000	Found	1863	26.2	-107.83333	(26.200000, -107.833330)
4	Willamette	24269	Valid	Iron, IIIAB	15500000	Found	1902	45.36667	-122.58333	(45.366670, -122.583330)
5	Morito	16745	Valid	Iron, IIIAB	10100000	Found	1600	27.05	-105.43333	(27.050000, -105.433330)
6	Brenham	5136	Valid	Pallasite, PMG-an	4300000	Found	1882	37.5825	-99.16361	(37.582500, -99.163610)
7	Old Woman	18007	Valid	Iron, IIAB	2753000	Found	1976	34.46667	-115.23333	(34.466670, -115.233330)
8	Navajo	16926	Valid	Iron, IIAB	2184000	Found	1921	35.33333	-109.5	(35.333330, -109.500000)
9	Coahuila	5387	Valid	Iron, IIAB	2100000	Found	1837	28.7	-102.73333	(28.700000, -102.733330)
10	Allende	2278	Valid	CV3	2000000	Fell	1969	26.96667	-105.31667	(26.966670, -105.316670)

Moreover, we go further into guessing the geographical most probable places to have meteorite landings in the US. These places are represented in a heat map in Figure 9.

Since we consider the US as a sample space of the entire map. We go further into one of the techniques regarding unsupervised learning datasets such as ours. We cluster the meteorite landings in the US and show them over the map to summarize the information in the heat map and provide useful knowledge for future astrophysicists and geologists. The clustering map is shown in Figure 10.

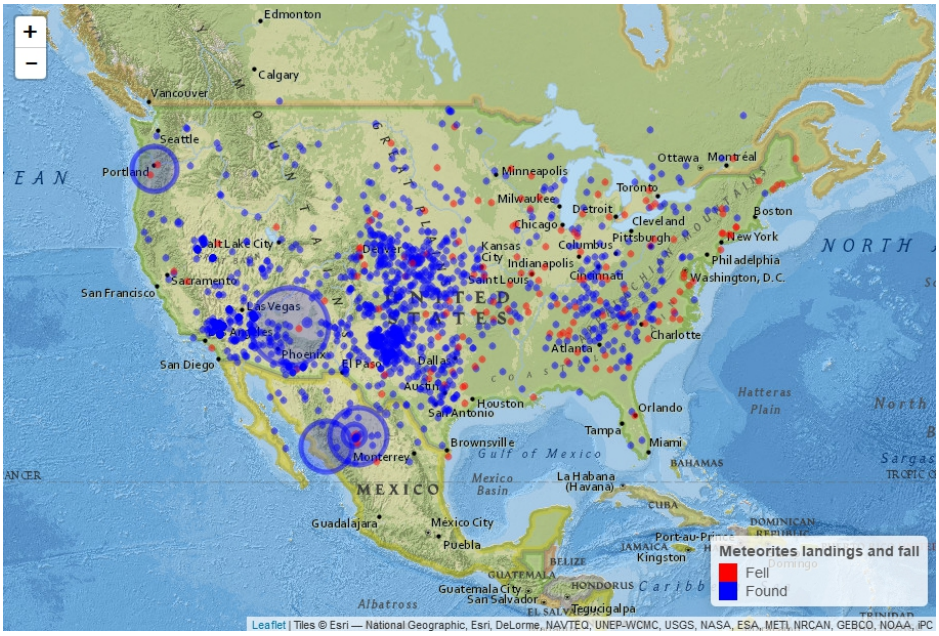

Figure 8: Map of Meteorites by Mass in US.

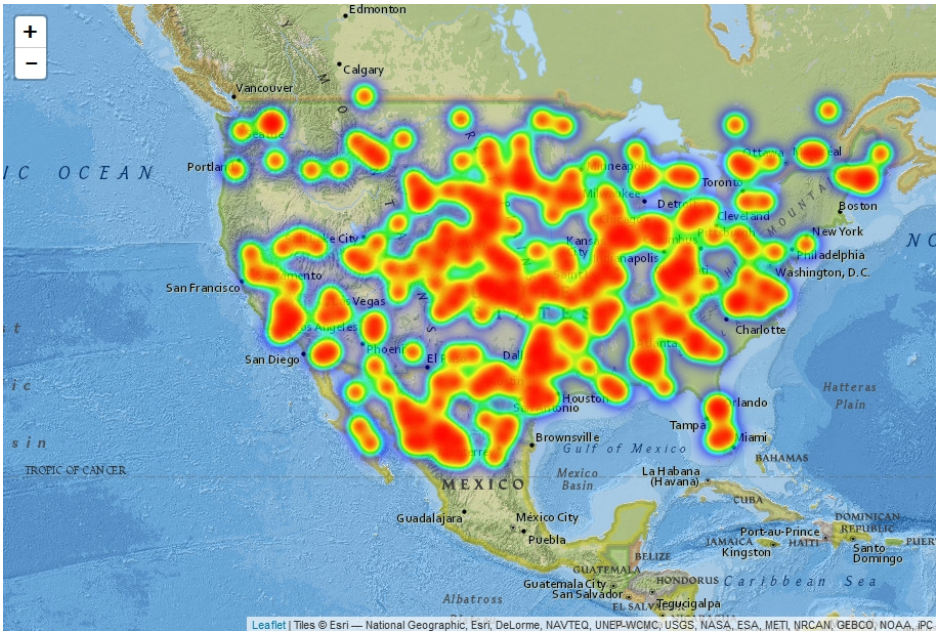

Figure 9: Heat map of the US.

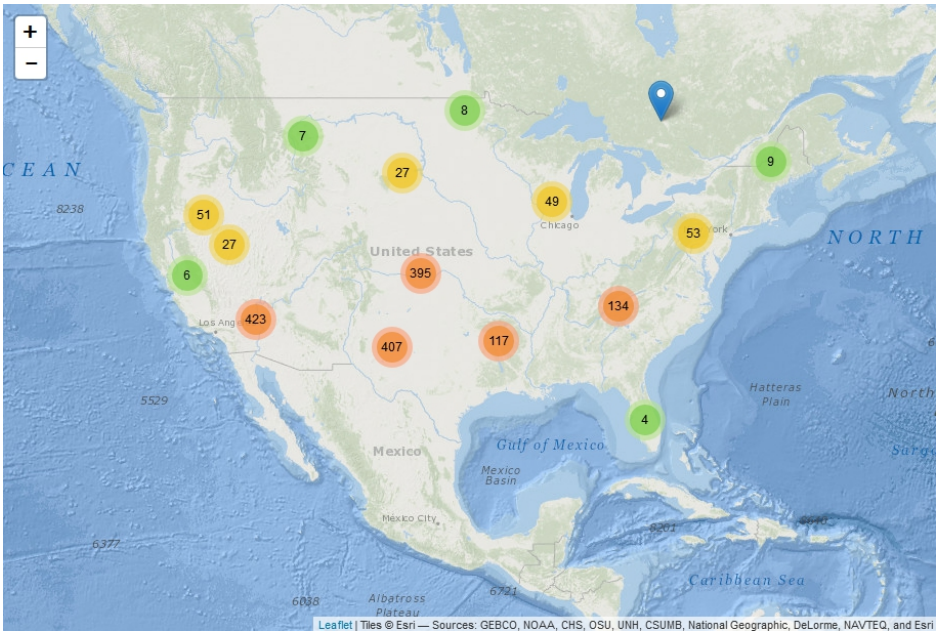

Figure 10: Clustering map of the US.

5.5 Year Analysis

One of the most popular misconceptions is that meteorite landings are growing bigger in numbers with time which is completely false according to our findings of the dataset and to our Figure 11 which indicates that there is absolutely no solid correlation between the count of meteorite landings and time.

6 Deployment

This work of analysis is implemented using R and R studio software. Our implementation is written in a single long script with proper commenting to explain the executed operations. Then, we stored the project on a GitHub repository on the the following link:
`https://github.com/Sabrout/MeteoriteAnalysis`

Surprisingly, it was expected to implement a machine learning algorithm to predict or solve a problem. On the other hand, such idea does not fit natural observations of phenomena in our case which leads us to do more

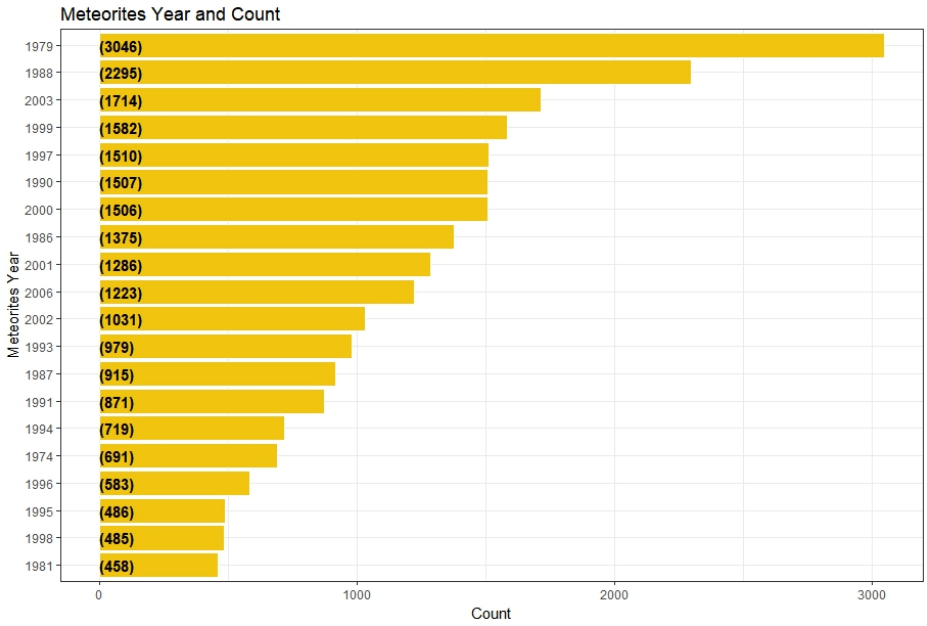

Figure 11: Meteorites Year and Count.

of a data analysis approach to find strong correlations more than a problem solving oriented approach.

7 Conclusion

To sum up, we discovered more knowledge about meteorite landings when we applied data mining techniques. Such knowledge might be a little bit vague at the current time being. Nonetheless, progressing on this analysis with deeper tools and analysis might lead to even more important and more interesting findings. This is considered a milestone for a promising future work.